初级保健领域内量表的设计与开发：实用步骤与统计方法

王飞，汤靖琪，孙小楠，等.初级保健领域内量表的设计与开发：实用步骤与统计方法［J］. 中国全科医学，2022.［Epub ahead of print］. DOI: 10.12114/j.issn.1007-9572.2022.0819

王飞[1]，汤靖琪[2]，孙小楠[3]，孙昕霙[4]，黎俊[5]，孟星星[2*]，吴一波[4*]

1.100875 北京市，北京师范大学认知神经科学与学习国家重点实验室

2.230039 合肥市，安徽大学哲学学院

3.150081 哈尔滨市，哈尔滨医科大学人文社会科学学院

4.100191 北京市，北京大学公共卫生学院

5.100191 北京市，北京大学第三医院全科医学科

*通讯作者：孟星星，讲师，E-mail：614997175@qq.com；吴一波，博士研究生；E-mail：bjmuwuyibo@outlook.com

【摘要】 本研究概述了在初级卫生保健领域内设计和开发有效可靠问卷的统计方法和实用步骤。回顾了一系列关于问卷编制和量表设计的研究，并制定了一套在初级保健领域内量表设计的标准流程。该流程涉及量表设计过程中关键的实用步骤以及统计学方法，并通过以往该领域内的相关研究案例加以说明。我们建议初级卫生保健问卷的七步编制方法如下：（1）定义测量的构想；（2）生成条目池；（3）选择评分系统和回答格式；（4）预测试（评估内容效度和表面效度等）；（5）通过项目分析剔除条目；（6）量表的初次评价，包括量表的信效度评价，以及因素分析或 Rasch 分析；（7）量表的再次评价，重新检验量表的性质，包括重测信度和结构效度。总的来说，量表设计类研究应严格按照量表编制的标准步骤，综合使用 Rasch 模型和因素分析的方法，将会使测量的结果更加客观。

【关键词】 初级保健；量表设计；因素分析；Rasch 模型

1.100875，北京市，北京师范大学认知神经科学与学习国家重点实验室 2.230039，合肥市，安徽大学哲学学院 3.150081，哈尔滨市，哈尔滨医科大学人文社会科学学院 4.100191，北京市，北京大学公共卫生学院 5.100191，北京市，北京大学第三医院

*通讯作者：孟星星，讲师，E-mail：614997175@qq.com；吴一波，博士研究生；E-mail：bjmuwuyibo@outlook.com

Design and development of scales within the primary care domain: practical steps and statistical methods

WANG Fei[1], TANG Jingqi[2], SUN Xiaonan[3], SUN Xinying[4], LI Jun[5], MENG Xingxing[2*],WU Yibo[4*]

1.State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University , Beijing 100875, China

2.School of Philosophy, Anhui University,Hefei 230039,China

3.School of Humanities and Social Sciences, Harbin Medical University, Harbin 150081, China

4.School of Public Health, Peking University,Beijing 100191,China

5.Peking University Third Hospital,Beijing 100191,China

**Corresponding author: MENG Xingxing, Lecturer; E-mail: 614997175@qq.com; WU Yibo, PhD student; E-mail: bjmuwuyibo@outlook.com*

【Abstract】 This study outlines statistical methods and practical steps for the design and development of valid and reliable questionnaires within the primary health care domain. A series of studies on questionnaire development and scale design are reviewed and a standard process for scale design within the primary care domain is developed. The process addresses key practical steps in the scale design process as well as statistical methods, and is illustrated by examples from previous relevant studies within the field. We suggest the following seven-step approach to primary health care questionnaire development: (1) defining the conceptions to be measured; (2) generating the pool of items; (3) selecting the scoring system and response format; (4) pretesting (assessing content validity and face validity, etc.); (5) eliminating items by item analysis; (6) initial evaluation of the scale, including reliability evaluation of the scale, and factor analysis or Rasch analysis; (7) re-evaluation of the scale, which reexamines the nature of the scale, including retest reliability and construct validity. In general, scale design type of studies should strictly

follow the standard steps of scale development, and the integrated use of Rasch model and factor analysis will make the results of measurement more objective.

【Key words】 primary care; scale devolpment; factor analysis; Rasch model

1 引言

世界卫生组织（WHO）在 1977 年第 30 届世界卫生大会上提出“人人享有健康”的宏伟目标，并指出初级卫生保健是实现这一目标的基本途径和关键[1]。全科医生作为初级保健服务的主要提供者，需要对来访者的特质做出准确的判断，才能给出更为合理的建议。而量表作为一种测量受测者某一特质的量具已被广泛运用于社会科学和医学当中，在初级保健领域内进行量表设计与开发有利于帮助研究者或全科医生测量出被试某一特质的程度。

然而，量表的设计与开发涉及到多个复杂且耗时的步骤，这些程序可能会令人望而却步，并且通常会忽略其中的部分程序[2]。这就造成了目前量表设计领域内问题的出现，如一项使用问卷评估运动员和教练的营养态度和营养知识的研究发现，大约 70%的纳入研究使用了效度和可靠性未知的工具，67% 使用了未经过验证的工具[3]。陈文雄编制的孤独症筛查量表中个别项目的信效度较差，但仍然保留在正式量表中[4]。这些未经信效度验证或信效度较差的量表会严重限制结论的得出，甚至会起到负面作用。因此，目前急需能够指导初级保健领域内量表设计研究的标准流程。除此之外，我们发现，初级保健领域内的量表设计研究绝大多数是在经典测量理论的框架之下进行的，这一技术对于量表心理测量学特性的验证是至关重要的，但由于经典测量理论的固有缺陷，往往不能保证测量的客观性。Rasch 模型的兴起为这一问题提供了很好的解决方式，Rasch 模型以自然科学领域内的客观测量当作标杆, 为社会科学领域内的测量建立起一套客观标准, 以确保测量所提供的信息更为客观和可靠[5]。

基于此，本研究将从经典测量理论和 Rasch 模型两个角度来总结目前国内外初级卫生保健领域内常用的问卷编制和量表设计方法，通过对具体步骤和统计方法的阐述帮助该领域内的研究者更好地开展研究。

2 实用步骤与统计方法

2.1 定义测量的构念

在初级保健领域内进行量表开发，其中最重要的一步就是对所需要测量的构念进行准确、概括的定义。定义中既需要解释所要测量构念的内涵和外延，还需要解释这

一构念的结构是什么。这种定义通常由经典教材、指南或该领域权威专家给出，也可以是基于大量文献和调查总结出来。前者在临床较为常用，为进一步扩展相关方法学应用，我们以基于大量调查和专家访谈确立定义为例。例如在 Wang 等人的研究中使用的就是 Weiss-Laxer 等人基于大量调查和专家访谈确立的定义：（1）研究者首先联系知名的家庭健康领域的研究人员组成专家小组，由研究执行者组成领导小组，共同明确了专家访谈的最终目标；（2）通过第一轮专家咨询，专家组提出并共同修改“家庭健康”的概念，由领导小组将概念划分为六个不同的领域；（3）专家进一步确认各个领域的内容和包含的概念，并按照重要性和可行性程度进行划分。最终得出家庭健康的定义：它是家庭单位层面的资源，从每个家庭成员的健康、他们的互动和能力，以及家庭的身体、社会、情感、经济和医疗资源的交叉点发展而来[6]，并在量表编制过程中选用重要的四个因素：家庭/社会/情感健康过程、家庭健康生活方式、家庭健康资源和家庭外部社会支持。Weiss-Laxer 等人在研究开始前界定了构念的内涵，其中包含了想要去测量的家庭健康的确切主题，同时也涵盖了家庭健康的相关维度，为研究的顺利推进奠定了基础，其方法值得研究人员学习。研究者也可以根据定义来确定问卷的初始维度和预期目的，使得初始测试尽可能多样化。

2.2 生成条目池

在完成测量构念的定义后，研究者就开始制作初始维度的条目池。代表同一维度的条目池要尽可能冗余，以确保最后能够符合预期条目，同时避免在后期数据处理过程中删减条目造成的条目数不够等问题。一般来说，研究者所编制量表的条目至少要达到最终保留版本的 2 倍。

条目池的生成通常是以经典教材、指南、文献和理论为指导，结合临床问题的前人研究或已有问卷，通过对已有资料的评估，编制出能够测量各维度特征的问题。因此，在编制量表条目池之前一定要明确各维度的定义，根据各个维度的定义来编制符合其含义的问题。例如在高志强等人编制成功恐惧问卷时，通过对已有研究的整理与分析，总结出了成功恐惧的结构维度分别是生活品质、家庭幸福、身体健康、心理健康、人际关系和恋爱择偶，再围绕这六个维度编制了最初的条目池并针对施测人群进行了初始化的结构化访谈和半开放式的问卷调查[7]。

在量表设计的语言方面也要遵循一定的原则，在编制量表条目时使用的语言应尽可能简单明了，避免使用专业性词汇和双重否定，因为它们常常会让受访者感到困惑；条目的语言尽量避免涉及社会禁忌和个人隐私，防止出现受访者的抵触情绪，干

扰研究；还要注意语言的使用一定要符合受测者所处地区的文化规范，必要时需进行调整。在成功恐惧问卷的最初编制中，在完成对量表内容的制定后还邀请中文系专家对量表语言进行评估，排除语意重复和歧义的条目，得到初始量表。

2.3 选择评分系统和响应格式

2.3.1 响应格式

响应格式选择通常与条目池的生成同步进行，研究者需要根据实际情况和调查的具体目的来选择适合该研究的评分系统和响应格式。

首先，研究者需要确定所编制条目池中每个问题的响应格式，是采用开放式提问的方式还是封闭式提问。开放式提问要求施测对象提供每个问题的答案，这对于受访者和研究人员来说更难回答，同时给出的答案往往也是多样的，不利于进行编码计分。开放式提问的好处是可以为研究者提供更多的思路，一般更适合在初始调查中使用，而在一个成型的量表中使用的并不多。因此，在初级保健领域内的研究中，使用较多的仍然是封闭式提问。封闭式提问会给出具体的选项，对施测对象来说更容易回答，但这也会造成其他的问题，如答案是设置单选还是多选？给出的可选择的答案不同是否会影响测量的结果呢？这在量表设计类研究中都是不可忽略的。

在绝大多数量表设计类研究中使用较多的是单选题，但是多项选择仍然是有价值的，因为很多时候一个问题并不会只有一个答案，而多项选择往往能够提供关于该问题更多的信息。孙昕霙等人（2022）利用项目反应理论开发出了评价糖尿病功能性健康素养量表，该量表一共包含 30 道题，其中有三道是多选题，它们提供了与糖尿病功能性健康素养有关的更多的信息。在评分方面，孙昕霙等人将多选题按选项数量每答对 1 个选项计相应分值，答不知道计 0 分[8]，但这种计分方式较为复杂，同时也会受到选项设置的干扰。一般来说，“选择所有正确的选项”的问题可能难以“编码”和评分，应尽可能避免[2]。此外，在封闭式提问设置选项时，仍然需要加以注意。例如在量表选项设置中是否应该加入“不确定”这一选项，Alsaffar 在翻译营养知识问卷时就使用了“不确定”这一选项[9]，但 Folasire 等人对此提出了质疑[10]，他们认为“不确定”选项容易导致那些对选项有很好了解的人在信心低下时避免回答或因为懒惰而选择逃避。除此之外，研究者还应避免将“其他”类别作为选项，当然只有在仔细确定了几乎所有可能存在的潜在类别之后，才能做出不提供“其他”选项的决定。

2.3.2 评分系统

在一份量表中，评分系统的选择往往需要结合具体的条目来进行设置。一般来说，

当问题回答有正误之分时，只需将正确的选择都记 1 分，而将错误的选择记 0 分，然而在绝大多数时候，受测对象很难做到绝对的二分，因此在实际研究中，最常用的评分系统是李克特式评分系统，如李克特式 5 点计分、7 点计分、9 点计分等。例如胡海利等人在编制中学生心理复原力量表时，便采用的五级计分法，以“从不”、“偶尔”、“有时”、“经常”和“总是”5 个等级进行程度评定，分别记为 1, 2, 3, 4 和 5 分[11]。而在涉及态度的研究中，研究者更倾向于使用“非常不同意”、“有些不同意”、“中立”、“有些同意”和“非常同意”5 个等级，计分仍然是从 1 到 5。这两者均属于李克特式五点计分，而七点计分和九点计分则是在五点的基础上进一步将选项细分。那么在研究中该如何选择李克特式量尺点数（如 5 点计分、7 点计分、9 点计分等）呢？Berdie（1986）认为当调查的对象具有较多的知识和较高的兴趣时，量表则需要更多的态度量尺点数，此时使用七点或九点计分比五点计分更为合适，因为当态度量尺点数越少，偏态程度越大[12]。

此外，在研究过程中，哪怕是收集了数据后，不同量尺点数的李克特式计分之间仍然可以转换。这种转换是通过 Rasch 模型来实现的，Rasch 模型可以系统地分析每个选项的测量特性，通过绘制选项概率曲线(Category Probability Curve, CPC)可以判断是否存在选项等级的滥用和缺失状况[13]。以 2021 年中国家庭健康指数中的法式烟草依赖评估量表（FTND）为例作图，FTND 的条目 1 内容如下：“您早晨醒后多长时间吸第一支烟？60 分钟（Category0），31-60 分钟（Category1），6-30 分钟（Category2），<=5 分钟（Category3）”。图 1 为条目 1 的选项概率曲线图，图中每条曲线对应一个选项，横轴代表被试烟草依赖的程度(从左往右递增)，纵轴代表被试选择的概率。以某位烟草依赖程度为-4 的被试为例,他选择“Category0”的概率约为 95%,选择“Category1”的概率约为 5%,选择其他选项的概率接近于 0。因此，该被试选择“Category0”的可能最大。以此类推，在 Category0 与 Category2 交点左侧，选择“Category0”的概率最大；在 Category0 与 Category2 交点和 Category2 与 Category3 交点之间，选择“Category2”的概率最大；在 Category2 与 Category3 交点右侧，选择“Category3”的概率最大。我们发现,测量过程中,“Category1”选项的使用率偏低,出现了李克特式等级滥用的情况。根据 Linacre 的建议，当出现李克特式等级滥用时，应考虑将相应的选项与相邻的选项进行合并[14]。因此，这里可以考虑将 Category1 与 Category2 合并为 6-60 分钟。但合并选项之后的量表仍需要再进行检验。

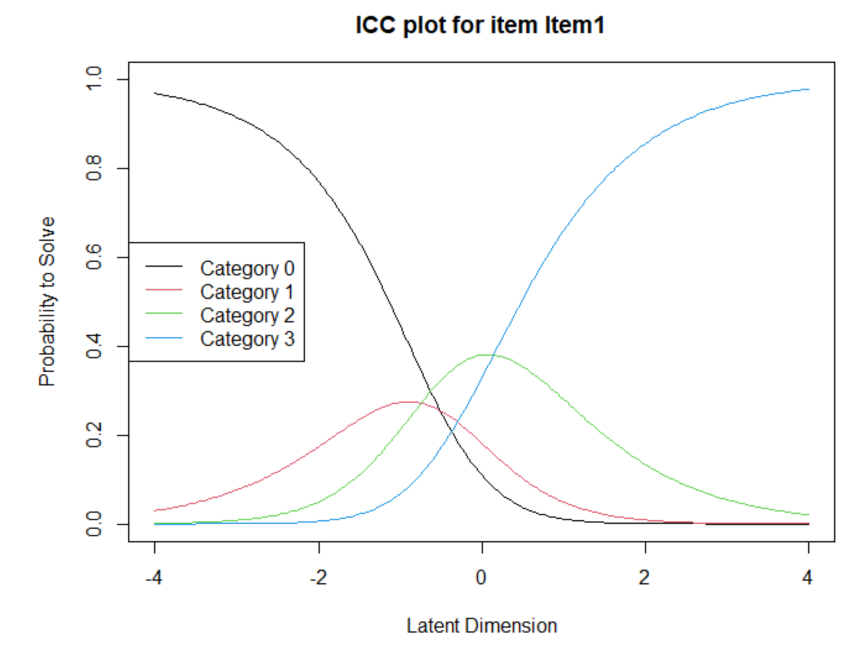


图 1 FTND 的条目 1 的选项概率曲线图

2.4 预测试

定性预测试是任何问卷或心理测量工具开发、翻译或修订的关键阶段。选取小样本受访人群进行小范围预测试，目的是验证目标受众人群是否理解条目问题与选项，从受访者角度评价条目表述是否有歧义，若出现语义理解困难、框架不清晰等问题，将修改条目后进行新一轮预测试直至确保所有受访者理解条目含义且内容可接受[15]。预测试主要采用便利抽样法，尽可能选择 30 份或以上样本以确保数据分析的稳定性与可靠性[15]，并对目标人群进行问卷填写感受与理解度调查。例如程彦如等人在编制失能老年人照顾者居家照护行为量表时，采用便利抽样选取某 3 个社区的 102 名失能老年人照顾者作为预测试对象。

预测试环节中需进行量表的表面效度的测评，即从受访对象角度看测评工具内容是否与测评目的一致，表面效度并不是真正的效度指标。在实际应用中，如果直接阅读问卷条目能够明显觉察问卷的测量意图，则该问卷表面效度较高。例如，测量护理人员洗手状况的问卷中，涉及洗手次数、时长以及方法等，所以此问卷具备表面效度[16]。在初级保健领域内，研究者想要考察患者关于保健领域内的行为情况或者针对某一病情进行详细询问，必然应当提高量表的表面效度，确保“所答即所问”；然而在涉及个人隐私方面或影响社会形象的问题上，表面效度过高可能会导致欺骗和隐瞒行为的出现，因此表面效度的设置需要依据具体研究目的而设定。

2.5 通过项目分析剔除条目

在初级保健领域内的量表编制过程中，应当在预测试实施完成后对其进行项目分析，该步骤为进一步修订量表提供依据，也是后续正确评价量表的前提。项目分析的实质是探究每个题项的差异，检验其质量，并依据一定的标准对其进行修订或剔除，保障项目之间的同质性与量表的可靠性。研究者主要可以从项目的难度、项目区分度和项目功能的差异三个方面考察。

2.5.1 项目难度

项目的难度是指在完成测验项目时所遇到的困难程度，是对测试者的作答情况进行评估的一个指标，作答正确率越高，难度越低。设置测验难度水平的目的是在于通过研究者开发的量表将不同的受测者尽可能区分开来，最大程度上体现受测者的差异，体现量表的鉴别力。正如步骤三所说，不同的量表类型适宜设置不同的计分系统，对于非二分法积分项目的难度可以采用所有受测者某一项目的平均得分与该题目满分之比来计算难度。比如在一项关于大学生健康素养的研究当中，研究者将多项选择题的反应进行重新编码，换算成另一种比例，对于正确值小于 0.2 或大于 0.8 的项目都进行了重评，并考虑是否删除[17]。过高或者过低的难度值都会给得分的分布和分数的离散程度带来影响，在实际操作过程中研究者应当考虑量表的性质和目的，科学设置合理的难度临界值。

Rasch 模型与经典测量理论所运用的方法有所不同，它主要强调了测量的客观性和可比性特征，因此对于测量难度这一指标该模型指出题目难度必须独立于样本被试分布，即抽样的人群在选择选项时时不受题目难度的影响，同时个体的能力也应当独立于测量题目的难度分布。即题目的难度不随着被试样本的变化而变化，不受被试能力水平高低的影响。因此 Rasch 测量能够提供关于个体能力和题目难度的等距分数，将个体能力水平和题目难度水平置于同一个 Logit 量尺中进行对比，刻画被试能力水平和项目难度水平的人-项目图（Person-Item Map），见图 2，图 2 是生活满意度量表的人-项目图，由该图可知，图中的黑点主要位于 0-2 之间，这意味在生活满意度量表项目中，中等及偏高水平生活满意度的被试者提供的信息量最大，但不适用于用来评定生活满意度水平较低的被试。不同的被试和项目就分布在这样一张图表中，可为研究者提供更多的信息。如果研究者计算出来的难度阈值和均值围绕在 0 附近，这就表明试题的难度适中，如惠建荣等人关于中风患者的生活质量量表的质量分析中，统计结果显示所有条目的难度阈值为-0.32~0.67(M=0.00，SD=0.34)[18]，这意味着所有条目的认可度都处于中等水平，较为良好。如果在量表开发过程当中项目难度水平过高或者过低，

则说明该题目所代表的行为或者维度出现频率并不高，或对于被试来说过难，而这样的量表往往只有在针对特定人群（过高或过低水平的被试）时准确度更高。

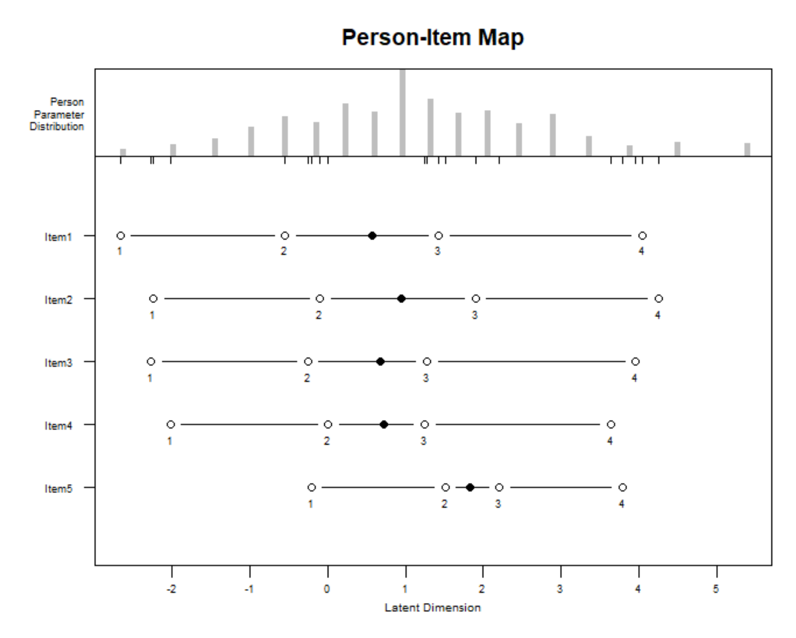


图 2 生活满意度量表的人-项目图

2.5.2 项目区分度

区分度的考察目的在于检验设计的量表是否真正能将两类不同的人区分开来，达到研究者预先的设想。主要包括鉴别指数法、相关法和 CITC 法。

鉴别指数的计算方法并不复杂，在统计好所有受测者的总分后按分数高低依次排序，测量学上一般以前后 27%的比例划分出高分组和低分组，对两组人各题的得分进行独立样本 T 检验，最终对于未表现出显著差异的题项单独考虑，必要情况下可以剔除以保障量表的准确性。或可以选择采用计算项目得分和测验总分的相关系数（PT-mesure）作为区分度指标，相关系数越大区分度则越高，最终综合考量是否剔除相关度不佳的项目；矫正项总计相关性（Corrected Item-Total Correlation, CITC）也可以用来考察量表维度中项目之间的相关性，如果大于 0.5 则说明该题项与其他项之间有着较高的相关，如果低于 0.5 则可以考虑删除该项目后观察 Cronbach α 系数的变化，或考虑修改该项目。如在花静等人对于儿童运动发育的研究中运用鉴别指数的方法测量各个项目之间，高分组低分组的得分差异，结果表示在 71 个条目上均有显著差异，因此在该阶段保留了所有条目[19]；而杨振等人在对老年健康促进量表进行信效度检验时，测量条目与量表总分的相关系数处于 0.406~0.752 之间[20]，呈中等程度相关（临界值为 0.3），随后结合信度系数对每个条目进行了进一步的检验。

在基于项目反应理论提出的 Rasch 模型当中，难度往往与区分度是密不可分的，

在中等难度下，项目的区分度往往最高。因此，项目的难度也可以通过人-项目图看出。图 2 中最下侧为 Rasch 标尺，从左到右测量值逐渐升高，对于每个被试而言，所处位置越靠近右端，说明生活满意度感受越高。图中条形高度表示位于这一位置被试的数量，被试分布越集中说明该量表的区分度越小，分布越分散说明量表的区分度越大。在图中我们可以看出在 5 个项目上，被试的掌握水平基本上都呈偏态分布，并集中分布在 0logit 到 2 logit 之间。这说明在 5 条项目中，该量表的区分度较差，在区分生活满意度较差的被试时较为困难。如赵福菓等人在编制奥尔维斯欺负量表时，使用 Rasch 模型发现难度分布非常集中，导致量表对不同霸凌/被霸凌程度被试的区分效果较差，尤其难以区分高霸凌/被霸凌群体[13]。

2.5.3 项目功能差异

项目功能差异（Differential Item Functioning，DIF）是指两组被试在某个项目上的表现差异，代表了项目对不同的被试有不同的统计特性，如果在同一项目上正确作答的概率不同，达到某一临界值，那么该项目则存在偏差，需要进一步的探究差异的来源[21]。基于项目反应理论的 Rasch 模型倾向于运用统计检验的方法计算 DIF，同时随着该理论模型影响力的进一步扩大，不同的学者提出了不同的计算方法。通过运用 Mantel-Haenszel（M-H 方法）检验法检验被试个人特征变量带来的 DIF，当差异大于 0.5 且 $p<0.05$ 时认为题目存在项目功能差异[22]；如杜海燕和李付鹏应用 M-H 方法进行 DIF 检验时发现第 9 题、第 39 题和第 58 题呈现出中等或较为严重的 DIF 现象[23]。也可以通过 Lord 卡方检验法、运用 R 语言软件进行项目功能差异检验，分析结果中 $X^2{}_{13}$ 为项目功能差异指标，某一项中 $X^2{}_{13}$ 的大于 0.05 说明存在 DIF[24]；如高爽和张向葵应用 Rasch 模型分析 Rosenberg 自尊量表时便是使用 Lord 卡方检验法，结果发现项目 1 和项目 5 存在功能差异，即在这两个项目上，性别差异导致自尊水平不同[25]。对于多级计分题也可以使用方差分析法进行检验，比如在世卫组织残疾评估计划的开展过程中，发现不同性别的群体之间项目难度不同，研究者采用方差分析，通过性别和其他有可能产生 DIF 的项目进行对比，从而找出不合适的项目，进行修改[26]。

值得注意的是，项目分析的三大方面并非要求在编制量表时全部使用，而是根据量表的特征加以选择——量表是单项选择还是多项选择？是二分法还是多级计分？开发的量表是什么性质的，等等。在项目分析过程中发现的问题项是否剔除也不能一概而论，简单的删除难度过大、区分度不良或拟合度不高的项目都并非值得提倡的做法，因为过于完美的模型难以真实存在，它只是一种理想性的假设与指导，应当结合多项

指标的综合情况进行考虑。

2.6 量表的初次评价

2.6.1 基于经典测量理论的初次评价

经典测量理论（Classical Test Theory, CTT）也被称作真分数理论，20 世纪 50 年代趋于完善。该理论认为测验得到的分数 X 是由真分数 T 和随机误差 E 所组成，即，X=T+E，误差 E 的平均数为零，T 和 E 之间的相关为零。并在此基础之上，建立了测验项目的测量学指标，如信度、效度、难度和区分度等，并以此筛选测验项目、建立题库和构制测验[27]。前文中已经对如何利用难度和区分度筛选测验项目做出了详细说明，本节将介绍如何运用经典测量理论来完成测验的初次评价，即进行探索性因素分析和信效度分析。

2.6.1.1 探索性因素分析

探索性因素分析（Exploratory factor analysis，EFA）作为一种经典测量理论技术，已经被广泛运用于初级保健领域内的量表设计与开发之中。探索性因素分析主要是通过数学的方法探索量表中的变量或因素，以此来确定量表的具体维度和每个项目归属于哪个维度。接下来，本文将详细介绍探索性因素分析的过程。我们认为探索性因素分析中应该包括以下 4 个关键步骤（见图 3）。

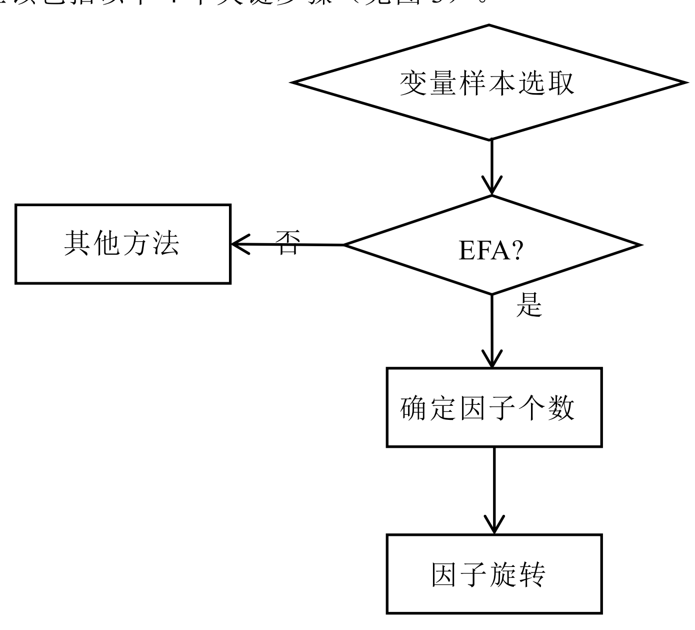


图 3 探索性因素分析流程图

（1）确定变量及样本

确定变量及样本是进行数据分析前的准备工作，这对于整个研究来说是至关重要

的。该阶段要求研究者根据以往研究和理论尽可能编制或收集与自己研究主题相关的条目，有时甚至需要包含一些与主题无关的条目。因为在经过探索性因素分析的筛选之后，剩下的条目往往会比原始条目少很多。如何决定条目的去留也是研究者需要关注的问题。常见的标准有因子载荷量、项目共同度、跨因子载荷等，通常认为成分矩阵中项目的因子载荷量>0.71 为优秀，>0.63 为非常好，>0.55 为好，>0.45 为一般，>0.32 为差[28]；项目共同度不能过低，一般认为项目共同度不得低于 0.30[29]；同一个项目不能在两个因子上都有着较高的载荷，如陈贵等人剔除了在不同因子上有相近的载荷且难以解释的项目[30]。在做因素分析之前，还需要注意样本量，因素分析的样本量不可太低，否则结果并没有太大的说服力，Corsuch 建议样本和变量比为 5：1，同时样本量不能低于 100。Nunnally 则推荐样本和变量比为 10：1[31]。

（2）确定是否可以进行探索性因素分析

探索性因素分析的目的是简化数据或者找出量表的基本数据结构，目前研究者普遍采用主成分分析法来进行探索性因素分析，因此在进行探索性因素分析之前需要确保因素分析的理论假设和统计假设得以满足。因素分析的理论假设认为这组变量中确实存在潜在结构，而统计假设要求观测变量之间存在较强的相关性。因此，在进行探索性因素分析前需要确保以下几个条件得以满足：项目间相关性大于 0.3、Bartlett 球形检验显著（$p<0.05$）以及抽样充分性 (MSA) 的 Kaiser-Meyer-Oklin (KMO) 度量至少为 0.6[2]。项目间相关性大于 0.3 要求研究者需要计算所有题目的相关性，如果所有或大部分相关小于 0.3 则不适合做探索性因素分析。球形检验和抽样充分性也是同样的道理，如郭静在修订中文版心理脆弱性问卷时进行了 Kaiser-Meyer-Oklin (KMO) 度量与 Bartlett 球形检验，结果显示 KMO=0.89，Bartlett 球形检验χ^2/df=25.31，$p<0.001$[32]。需要注意的是，这些参数合格仅代表可以进行因素分析而不是说明因素分析结果较好。

（3）确定因子个数

确定所选变量的因子结构，保留多少个因子是探索性因素分析中非常关键的一步，抽取过少或过多都会造成一定的问题，但实证研究中更倾向于保留较多的因子，因为抽取过度相比于抽取不足的因子载荷估计更加准确。因此研究者提出了多种检验方法来帮助决策，主要包括以下三种：□特征值大于 1，特征值大于 1 也叫 K1 原则，是研究者最常采用的标准之一。□解释方差总量，方差解释量也是基于主成分分析法的思想发展而来。关于因子解释多少总体方差合适并没有统一的标准，有研究者认为因子解释的方差总量应不得低于 50%[33]。例如，表 1 显示了 8 条目一般自我效能感量表的因

子分析结果，图中仅有一个特征值大于 1（5.753>1），研究者据此可以认为一般自我效能感量表是个单维度的量表，仅包含一个因子；不仅如此，表中还显示了该因子的方差解释量（71.91%>50%），这意味着该因子能够解释一般自我效能感 71.91%的变异，能较好地反映一般自我效能感。□碎石图，碎石图提供了因子数和特征值大小的图形表示，研究者只需要根据 EFA 给出的碎石图选择出现拐点时对应的因子数即可，这种方法简单方便，也更加直观。图 4 为一般自我效能感的碎石图，由图可知，在从第一个成分开始，特征值产生了巨大转折，因此可将第一个成分视为拐点，认为该量表仅包含一个因子。

表 1 一般自我效能感解释的总方差

成份	初始特征值			提取平方和载入		
	合计	方差的 %	累积 %	合计	方差的 %	累积 %
1	5.753	71.910	71.910	5.753	71.910	71.910
2	.515	6.441	78.351			
3	.388	4.845	83.196			
4	.306	3.829	87.024			
5	.295	3.683	90.708			
6	.276	3.444	94.151			
7	.244	3.055	97.206			
8	.224	2.794	100.000			

提取方法：主成份分析。

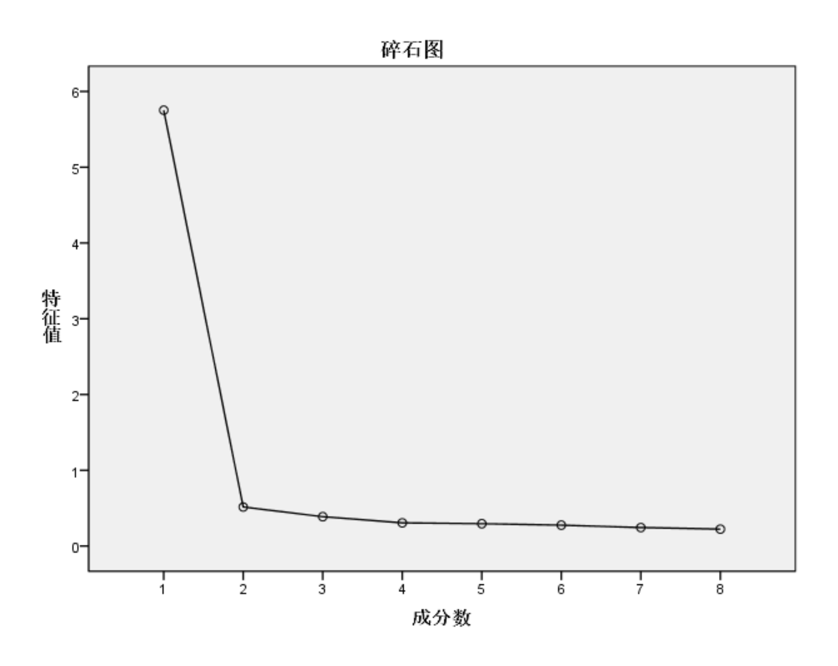


图 4 一般自我效能感量表的碎石图

（4）因子旋转

在确定了因子个数后，下一步就需要确定因子旋转的方法。因子旋转的方法可分为两大类：斜交旋转（Oblique Rotation）和正交旋转（Orthogonal Rotation）。与斜交旋转不同的是，正交旋转需要假设因子之间无相关，而斜交旋转则并不存在。就初级保健领域内的实证研究而言，因子之间往往都存在着或大或小的相关性，因此采用斜交旋转更加客观，然而目前已发表的绝大多数研究使用的多是正交旋转，它的结果更有利于研究者对因子结构做出解读，但这也容易对研究结论造成误导，因此，我们认为未来的研究者应该先用斜交旋转，如果发现因子见相关较小或没有相关时再考虑采用正交旋转。表 2 显示的是应用 promax 斜交旋转法进行因子旋转的结果，结果显示，该量表包含 2 个因子，其中 J1、J2、J3、J4、J5、J7、J8 属于因子 1，而 J6、J9、J10 属于因子 2。

表 2 promax 斜交旋转的结构矩阵

	成份	
	1	2
J1	.889	.145
J2	.879	.144
J3	.897	.178
J4	.895	.146
J5	.899	.172
J6	.177	.772
J7	.736	.011
J8	.704	.083
J9	.044	.835
J10	.141	.886

2.6.1.2 信度分析

经历了探索性因素分析的剔除条目后，正式量表已经成型，此时还需要利用该数据检验正式量表的信度。信度是指测量结果的稳定性。如果一个人的同一种特质能够用同一种测量工具反复测量，那么各种测量相互间的吻合程度就称为信度，有时也称为测量的可靠性。在经典测量理论中，衡量信度方法通常包括复本信度、重测信度、同质性信度、分半信度、和评分者信度。在临床研究中，由于复本信度较难获得，因此研究当中很少使用这一指标，研究者更倾向于使用重测信度、分半信度和同质性信度。

（1）重测信度

在量表设计类研究中，量表的跨时间一致性是一个衡量测量工具可靠性的重要指标。因此，在初级保健领域内进行量表的开发和设计时，需要报告该量表两次对同一组被试施测所得结果的一致性程度，其大小可用前后两次相同测验的皮尔逊积差相关系数来表示。如刘蕾等人在编制中文版老年人锻炼心理需求满足量表时报告了该量表的重测信度为 0.883，3 个维度的重测信度系数分别在 0.829~0.876 之间[34]。对于测验中的重测信度，一般公认的评价标准是：0.65~0.70，最小可接受值；0.70~0.80，相当好；0.80~0.90，非常好[35]。因此，刘蕾等人所编制量表的重测信度较好。但刘蕾等人并未报告两次施测的间隔，这也是影响重测信度的重要因素，在今后的研究中应该要加以注意，因为随着第二次测量的时间不同，它可以有不同的重测信度，

（2）复本信度

通过设计两个平行测验来测量同一批被试，所得结果的一致性程度则称为复本信度，它的大小可使用两个复本测验上同一批人测试的皮尔逊积差相关系数来表示。复本信度也是衡量量表可靠性一个指标，但是由于设计复本测验费时费力，同时又很难保证两个测验在内容和结果上一致，因此，其在测量领域内却并没有得到了广泛的使用。刘爱梅和刘院斌在编制适用于突发性耳聋患者的健康知信行问卷时就使用了这一信度，复本测验采用的是采用内容、应答形式相似的问卷进行调查，结果发现健康相关知识部分的复本信度为 0.88[36]，而复本信度的评价标准与重测信度基本上保持一致[35]，因此，该量表的复本信度较好。

（3）分半信度

也叫内部一致性系数，研究者需要将一个完整的测试分成对等的两半，比较参与测验的被试在新得到的两组上测验分数的一致性。分半信度是目前研究中使用最多的信度之一，研究者只需要在统计软件 SPSS 内进行简单操作即可算出该量表的分半信度。

（4）同质性信度

研究者可通过测量测验内部所有题项彼此之间的一致性程度得到同质性信度，即内部一致性系数。研究者一般采用 Cronbach’ s alpha 系数来衡量一个测验的内部一致性。alpha 系数是目前研究中使用最多的信度，与分半信度类似，研究者只需要在统计软件 SPSS 内进行简单操作即可算出该量表的 alpha 系数。吴明隆指出 alpha 系数最好在 0.80 以上，0.70~0.80 是可以接受的范围；分量表最好在 0.70 以上，0.60~0.70 是可以接受的范围[35]。

（5）评分者信度

由多个评分者给同一批人的答卷进行打分，通过计算得分的一致性，可以得到量表的评分者信度。其大小等于一个评分者的一组评分与另一个评分者的一组评分的肯德尔和谐系数。肯德尔和谐系数是表示多列等级数据相关程度的一种量数，常用于评价多个主评的评分一致性。

2.6.1.3 效度分析

在进行初级保健领域内开展量表设计研究时，还应检验所编制测验的效度。效度是一个测试或量表能够测量它试图测量的特征的程度。效度的理论定义是指在与测量目的相关的一系列测量中，真实变化(被测量变化引起的有效变化)与总变化(真实变化)的比值。测试效度可分为内容效度、结构效度和经验效度。

（1）内容效度

内容效度是由相关专家对测评工具的条目与内容范围的吻合度进行详尽、系统判断。其中参评专家的资质、专业范围是内容效度评估质量的基本保障，比如崔楚云等人选择 6 名护理领域专家（来自学校和医院的护理学教授、护理部主任以及临床护理专家）对量表内容效度进行评价，因为选择研究领域的教授或临床专家是开展内容效度评价是最常见的选择[37]。另外，内容效度在条目筛选中的定量评估包括多种指标计算，其中内容效度指数 (content validity index， CVI)由于计算简单，易于理解和交流，可对随机一致性进行校正等优点得到广泛应用：项目水平的内容效度指数 (I-CVI)可以评估各个项目的内容效度；量表层面的内容效度指数(S-CVI)用于衡量整个量表的内容效度。。例如，在完成冠心病病人二级预防服药依从性问卷的初步编制后，研究者依照 Likert 4 级评分法编制专家评定表，选项设定为不相关、修改否则不相关、很相关但仍需修改、十分相关四级，依次计为 1—4 分，发放给专家作答，回收后计算得出 I - CVI 和 S - CVI 均为 1.00[38]，表明内容效度良好。

（2）结构效度

测验在实际上所测到想要测量的理论和特质的程度即为量表的结构效度，它表示了一份量表在多大程度上能够说明测验理论的某种结构或特质。在实证研究中，研究者一般可以通过项目分析、探索性因子分析以及验证性因子分析（Confirmatory factor analysis，CFA）来衡量一个量表的结构效度。项目分析是通过计算量表各条目与所在维度的相关矩阵以及各维度之间的相关矩阵来检验量表各维度之间的关联性与独立性。如杨丽等人在认知风格问卷中使用了项目分析来衡量量表的结构效度，结果显示项目

与所在维度的相关得分均在 0.55 以上，基本分布在 0.56 到 0.75 之间，问卷的项目区分度良好，认知风格问卷四个维度之间存在中等相关，说明四个维度相互关联，同时相对独立[39]。探索性因素分析与上节所述基本一致，只不过这次不需要删减条目，一般来说，经历过探索性因素分析形成的问卷在检验其结构效度时应重新收取新的数据，对新的数据采用探索性因素分析或验证性因素分析来衡量。如吴一波等人在检验中文版杜克抗凝满意度量表 (DASS)的信效度时使用 AMOS 软件进行验证性因素分析来检验模型拟合，结果发现各项指标均显示四因素的 DASS 模型拟合良好（CMIN/DF = 1.825 < 5，GFI = 0.854 > 0.85，CFI = 0.938 > 0.9，RMSEA = 0.066 < 0.08，NFI = 0.875 < 0.9，TLI = 0.921 > 0.9），量表具备良好的结构效度[40]。

（3）实证效度

如果一个测验能够对处于具体情境中的被试的行为进行有效的估计，则称该测验具有良好的实证效度或校标关联效度。效标效度主要可以通过以下相关法、区分法和命中率法来进行衡量，而目前初级保健领域内的量表设计研究中使用较多的仍是相关法。相关法是测试成绩与效度变量之间的相关程度。计算出的相关系数为效度系数，效度系数的平方为效度。如游永恒等人就选取总体幸福感量表（GWB）作为效标来验证 Beck 抑郁量表的同时效度，再发放抑郁量表时同时要求作答校标量表，结果发现总体幸福感各个维度及总分与抑郁总分均有显著的相关性（P<0.001），这表明 BDI 量表具有较好的效标效度[41]。

2.6.2 基于 Rasch 模型的初次评价

拉希（Rasch）模型是一种基本特征模型，它通过个体在某项上的表现来衡量基本特征。拉希（Rasch）模型的基本原理是，一个人在具体题目上的具体表现是由这个人的能力和题目的难度来衡量的，因此个体反应的好坏完全取决于个体能力和项目难度。Rasch 模型是一种理想化的数学模型，因此 Rasch 模型对客观测量提出了两个要求：（1）对任何题目，能力高的个体应该比能力低的个体有更大可能做出正确回答；（2）任何个体在容易题目上的表现的更好，困难题目上表现更差[42]。尽管 Rasch 模型已经发展了数十年时间，但该模型仍然没有引起足够的重视，尤其是在初级保健领域。在“中国知网”(1915 至 2022 年)以”Rasch”为主题进行检索，结果只发现了核心期刊 160 篇，其中近五年（2017-2021）的研究占比高达 46.25%，这意味着近年来，Rasch 模型已渐渐被更多的研究者注意，然而这些研究仍然主要集中于心理学、教育学领域，涉及初级保健的文章仅有寥寥数篇。因此，在初级保健领域内开展 Rasch 模型研究非常

有必要。

2.6.2.1 单维性检验

项目反应理论(IRT)是一种关于个体回答问题的概率与潜在特质之间关系的数学表述，是区别于 CTT 的又一测量领域的经典理论。常见的 IRT 模型包括单参数模型、双参数模型和三参数模型[42]。Rasch 模型作为 IRT 单参数模型的一个特例，它的使用有一个前提，那就是量表具有单维性。单维性是指测量过程中有且仅有一种潜在特质影响被试作答。在这里需要注意的是，一种潜在特质并不意味着该量表只能有一个维度，只要量表中的各个维度都指向同一种特质即可。如陈圆圆等人在汉化营养素养评价工具发现该工具包含 6 个分量表，但分量表中包含的条目都要指向营养素养这一特质，于是她们针对分量表和全量表均做了 Rasch 分析[43]。一般采用 Rasch 模型残差主成分分析法（PCA）检验量表单维性，根据 Raiche 的建议，首因子残差标准化特征值在（1.4～2.1）之间即可认为该数据满足单维性的要求，适合 Rasch 模型[43]。如陈圆圆等人在进行汉化营养素养评价工具过程中进行单维性检验发现分量表 1～6 的首成分残差特征值分别介于 1.6-1.8 之间，总量表的首成分残差特征值是 3.1，这意味着该量表适合进行 Rasch 分析[43]。

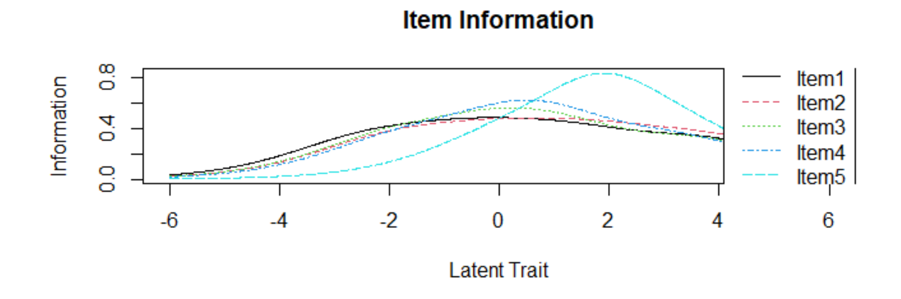
2.6.2.2 模型拟合度

从怀特图中，我们得知 Rasch 模型能够估计项目的难度和被试的能力水平，通过将实际的观测分数与每个被试在每个项目上答对的理论概率进行比较，即可评估 Rasch 模型的拟合情况。Rasch 模型通常需要计算两个拟合指标：加权均方拟合统计量(Outfit Mean Square,Infit MNSQ)和非加权均方拟合统计量(Outfit Mean Square,Infit MNSQ)，Infit MNSQ 与 Outfit MNSQ 接近于 1 表示模型拟合效果好。一般认为，当数据拟合良好时，Outfit 和 Infit 的 MNSQ 在 0.5 ~ 1.5 之间[44]。以生活满意度量表为例，我们收集了 569 份数据，使用 R 进行模型拟合度检验，结果见表 3。由表 3 可知，所有项目的参数基本都在可接受范围内，说明数据与模型达到了很好的拟合。题目 5（如果我能重新活过，差不多没有东西我想改变，1=不同意，2=有些不同意，3=中立，4=有些同意，5=同意）的 Outfit MNSQ 和 Infit MNSQ 参数值分别为 1.52 和 1.40，项目 5 的 Infit MNSQ 和 outfit MNSQ 参数值均大于 1.0，这意味着有较高生活满意度的人选择了低分，即不同意和有些不同意；而有着较低生活满意度的人选择了高分，即同意和有些同意。因此，题目 5 在区分被试生活满意度时误差较大，需要进一步考虑是否需要保留该条目。

表 3 生活满意度量表的模型拟合参数

	Chisq	df	p-value	Outfit MNSQ	Infit MNSQ	Outfit t	Infit t
Item1	343.86	538	1	0.64	0.63	-6.99	-7.10
Item2	324.13	538	1	0.60	0.60	-7.89	-7.86
Item3	307.71	538	1	0.57	0.59	-8.59	-8.26
Item4	496.31	538	0.90	0.92	0.91	-1.36	-1.54
Item5	817.96	538	0	1.52	1.4	7.21	5.90

此外，一个较好的项目或量表应该能够为测试提供较多的信息，降低对被试特质水平估计方面的误差。项目反应理论认为，用与被试特质水平相当的量表进行测试时，量表才能提供最精准的测量结果。在研究中，一般采用测试信息曲线进行测量，它可以反映当不同特征水平的被试完成完整量表的所有项目时，量表整体能提供准确评价的程度。其中项目的难度可参见横坐标，代表了被试的特质水平，每个刻度代表一个logit 单位，纵坐标代表信息量，即 Fisher 信息函数[13]。图 5 是生活满意度量表的测验信息曲线图，其中上半图是各个条目的测验信息曲线，下半图是总量表的测验信息曲线。总体而言，该量表在生活满意度估计值在 0～2 之间时准确率最高，能为中、高生活满意度的被试提供最大的信息。例如，高爽和张向葵在计算 Fisher 信息函数后发现，自尊的估计值在 0～-2 之间，可以提供最高的测量精度，为中、低自尊被试提供最多的信息[25]。



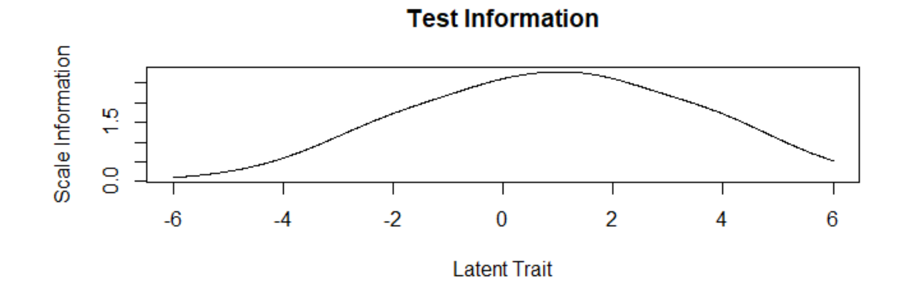


图 5 生活满意度的测验信息曲线

2.6.2.3 信度

Rasch 模型以分隔信度 (Person separation reliability，PSR) 衡量量表信度，分隔信

度可以通过计算个体所产生“真实”变异与总变异的比例得出，通常用于考察受试者在项目评定上的可靠性程度[13]。Rasch 模型测量的总体信度是通过计算个体水平上的解释率得到的，其值从 0 到 1。一般情况下，可靠性指标在 0.7 以上为可接受，0.8 以上为良好[5]。经计算获得，生活满意度量表的信度值为 0.80，信度较好。

2.7 量表的再次评价

从第一步到第六步，一个量表基本上已经成型。但由于量表条目的筛选和信效度检验均是采用同一份样本，该量表是否具有跨样本和跨时间的一致性仍然是未知的。因此，研究者应该使用正式量表重新收集一个新的样本，检验该量表在新样本上的信度与效度。当然，需要注意的是，如果研究者需要检验该量表的重测信度，那么第二批量表的被试中就应该包含一部分第一批施测的对象。由于信效度分析的相关内容已经在前一节中阐述过，研究者只需要使用相同的方法再次检验即可，便不过多赘述。这里仅对经典测量理论中使用验证性因素分析检验量表结构效度的方法进行阐述。

验证性因素分析是指在明确了观测指标和潜在因子之间隶属关系的前提下进行的假设检验，是理论驱动型分析。在经历了探索性因素分析以后，我们明确了正式量表的因子结构，因此，可以利用新数据构建验证性因素分析模型来检验量表的结构效度。再根据输出结果的拟合状况考虑是否需要进行模型修正。主要选用的拟合指标包含卡方自由度比值（$\chi2/df$）、适配度指数（GFI）和调整拟合优度指数（AGFI）、渐进残差均方和平方根（RMSEA）以及标准拟合指数（NFI）、增量拟合指数（IFI）、相对拟合指数（RFI）、相对拟合指数（CFI）、Tucker- Lewis 指数（TLI）等。这些参数的适配标准为:$\chi2/df<2$ 时（也有研究者认为 $\chi2/df<3$），表示假设模型的适配度较佳[45]。RMSEA<0.08，意味着模型尚可接受[6]。AGFI 与 GFI 均应>0.90，表示模型与数据有着良好的匹配度[46]。NFI、RFI、IFI、TLI、CFI 均应大于 0.90[46]。如果这些拟合值数未达到较好的适配标准，研究者则应考虑对模型进行修正，具体做法是利用 AMOS 报表呈现的 MI 值，释放两个测验误差变量彼此之间的关系，即在其之间建立共变关系[46]，从而达到对优化模型的目的。

3 讨论

量表设计类方法在初级保健领域内得到了充分地运用，这主要体现在量表设计研究的使用广度上。目前绝大多数研究中都涉及量表的使用，因此，一个量表的设计与开发是否合理便决定了该研究是否可靠。而目前关于量表设计研究仍存在较多不规范的地方，如信效度较差、缺乏关键步骤、统计错误等问题。总的来说，在初级保健领

域内开展量表设计类研究需要严格按照上述标准流程进行，这在一定程度上能够解决研究过程中步骤和统计方法使用不规范的情况。当然，为了更好地掌握这种方法，有些必需技能也是需要注意的。

量表设计类研究所需要的必要技能主要包括理论指导和统计检验。理论指导是自上而下的加工，是理论驱动的过程。理论指导要求研究者在开发量表前期和中期一定要阅读大量相关文献，了解所需要测量特质的结构以及现有理论和量表，只有在这些成熟的前人经验的基础之上，才能尽可能确保所编制量表的有效性。而统计检验是自下而上的加工，是数据驱动的过程。统计检验可以帮助研究者更好地发现项目编制过程中存在的问题，同时也是研究者筛选不好条目的重要参照。研究者通过统计学来检验量表的信度和效度，以此来保证这一量具的客观与有效。综上，理论指导和统计检验是量表设计类研究中两项必需的技能，只有将这两者很好地结合起来，从自下而上的自上而下的角度一起考虑，才能在最大程度上保证所设计测量工具的可靠性。

本研究较为系统地阐述了如何在初级保健领域内开展量表设计，但由于篇幅和专业性的限制，使得一部分的临床医生可能很难理解文中出现的专业术语，不仅如此，可能对于大多数全科医生来讲，如何选取一个合适的量表比设计一个量表更为直接有效。为此，我们在附件中提供了文中出现的一些专业词汇的解释以及全科医生应该如何选取量表的相关建议。此外，本研究还为研究者提供了继续深入学习量表设计类方法的学习资源清单，见表 4。总的来说，研究者在开展量表设计时需要严格遵守标准流程，在具体步骤中可参照清单中的相关资料，这样就能确保所设计量表的客观有效。

表 4 继续学习资源清单

序号	作者	名称	出版社/杂志	种类
1	陈新丰	R 语言-量表编制、统计分析与试题反应理论	东北财经大学出版社	书籍
2	罗伯特·F.德威利斯	量表编制：理论与应用	重庆大学出版社	书籍
3	王媛媛	医学量表的编制与评价：理论、方法与实例操作	北京大学医学出版社	书籍
4	简小珠 & 戴布云	SPSS23.0 统计分析在心理学与教育学中的应用	北京师范大学出版社	书籍
5	吴明隆	结构方程模型——AMOS 的操作与应用	重庆大学出版社	书籍
6	王孟成	潜变量建模与 Mplus 应用	重庆大学出版社	书籍
7	Kline, & Paul.	The Handbook of Psychological Testing	Routledge	书籍

8	Robert B. Frary	A Brief Guide to Questionnaire Development	Office of Measurement and Research Service Virginia Polytechnic Institute and State University	文章
9	R. Noah Padgett & Grant B. Morgan	Using the eRm Package for Rasch Modeling	Measurement: Interdisciplinary Research and Perspectives	文章
10	Baghaei, Purya&Doebler, Philipp	Introduction to the Rasch Poisson Counts Model: An R Tutorial	Psychological Reports	文章

4 结论

总之，我们为有兴趣在初级保健领域内开发或设计量表的研究人员概述了实用步骤与统计方法。我们建议所有在初级保健领域内进行量表设计时都应考虑本综述中描述的方法，研究者应严格按照量表编制的标准步骤，综合使用 Rasch 模型和因素分析的方法，将会使测量的结果更加客观。

作者贡献 王飞提出研究选题方向，负责数据处理，并撰写论文初稿；汤靖琪参与了论文初稿的撰写并进行了数据管理；孙小楠负责论文的修订；孙昕霙对文章提出了批判性的建议；黎俊则从全科医生的视角对文章进行了修改和完善；孟星星和吴一波全程指导了该研究，并负责文章的质量控制及审校，对文章整体负责；所有作者确认了论文的最终稿。

利益冲突情况 本文无利益冲突

致谢 我们要感谢安徽大学哲学学院的高志强副教授在心理测量领域给予的指导，正是因为高志强副教授的心理测量课程才让我们很早地就了解到了这一领域。还要感谢参与 2021 年家庭健康指数调查的全体调查员，正是因为有了你们的参与，才能有如此多的数据来支持文中的相关图表。

参考文献

[1] 王荣华, 王素平. 全科医生在我国医疗卫生服务体系中的作用研究 [J]. 中国全科医学, 2020, 23(04): 388-394+402.

[2] TRAKMAN G L, FORSYTH A, HOYE R, et al. Developing and validating a nutrition knowledge questionnaire: key methods and considerations [J]. Public Health Nutrition, 2017, 20(15): 2670-2679.

[3] KOUVELIOTI R, VAGENAS G. Methodological and statistical quality in research evaluating nutritional attitudes in sports [J]. International journal of sport nutrition and exercise metabolism, 2015, 25: 624-635.
[4] 金映彤, 陈苏琴, 包韵歆, 等. 儿童孤独症谱系障碍评估工具的研究进展 [J]. 护理实践与研究, 2021, 18(09): 1325-1329.
[5] BOND T G, FOX C M. Applying the Rasch model.Fundamental measurement in the human sciences(3rd ed.) [M]. New York: NY.Routledge, 2015.
[6] WANG F, WU Y, SUN X, et al. Reliability and validity of the Chinese version of a short form of the family health scale [J]. Bmc Primary Care, 2022, 23(1).
[7] 高志强, 张腾霄. 成功恐惧问卷的编制及应用 [J]. 中国临床心理学杂志, 2011, 19(05): 602-605+86.
[8] 孙昕霙, 朱小柔, 巩俐彤. 基于项目反应理论的糖尿病功能性健康素养量表评价 [J]. 中国健康教育, 2022, 38(01): 18-22.
[9] ALSAFFAR A A. Validation of a general nutrition knowledge questionnaire in a Turkish student sample [J]. Public Health Nutrition, 2012, 15(11): 2074-2085.
[10] FOLASIRE O F, AKOMOLAFE A A, SANUSI R A. Does Nutrition Knowledge and Practice of Athletes Translate to Enhanced Athletic Performance? Cross-Sectional Study Amongst Nigerian Undergraduate Athletes [J]. Global journal of health science, 2015, 7(5): 215-225.
[11] 胡海利, 张洪波, 王君, 等. 中学生心理复原力量表的编制及其初步评价 [J]. 中国学校卫生, 2009, 30(12): 1097-1099.
[12] BERDIE D R. The optimum number of surveg research scale points:what respondents sau.the meeting of the American EducationalResearch Association[C].San Francisco: CA.F, 1986.
[13] 赵福菓, 何壮, 袁淑莉, 等. 奥尔维斯欺负量表的 Rasch 模型分析 [J]. 西南大学学报(社会科学版), 2020, 46(05): 115-121.
[14] LINACRE J M. Optimizing rating scale category effectiveness [J]. Journal of applied measurement, 2002, 3(1): 85-106.
[15] PERNEGER T V, COURVOISIER D S, HUDELSON P M, et al. Sample size for pre-tests of questionnaires [J]. Quality of Life Research, 2015, 24(1): 147-151.
[16] 表面效度 [J]. 中国护理管理, 2016, 16(07): 905.
[17] RABIN L A, MILES R T, KAMATA A, et al. Development, item analysis, and initial reliability and validity of three forms of a multiple-choice mental health literacy assessment for college students (MHLA-c) [J]. Psychiatry Research, 2021, 300.
[18] 惠建荣, 裴建, 王院春, 等. 针刺干预中风专用生活质量量表的 Rasch 分析 [J]. 中国针灸, 2013, 33(4): 363-366.
[19] 花静, 张郦君, 古桂雄, 等. 城市学龄前儿童运动发育家庭环境量表的初步编制 [J]. 中国学校卫生, 2011, 32(2): 161-163.
[20] 杨振, 张会君. 老年健康促进量表的跨文化调适及信效度检验 [J]. 护理学杂志, 2021, 36(19): 91-94.
[21] LAI J S, CELLA D, CHANG C H, et al. Item banking to improve, shorten and computerize self-reported fatigue: An illustration of steps to create a core item bank from the FACIT-Fatigue Scale [J]. Quality of Life Research, 2003, 12(5): 485-501.

[22] ZWICK R, THAYER D T, LEWIS C. An Empirical Bayes Approach to Mantel-Haenszel DIF Analysis [J]. Journal of Educational Measurement, 1999, 36(1): 1-28.
[23] 杜海燕，李付鹏. 样本容量对 Mantel-Haenszel 方法检验 DIF 效应的影响分析 [J]. 考试研究, 2016, (05): 55-62.
[24] CHOI S W, GIBBONS L E, CRANE P K. lordif: An R Package fo rDetecting Differential Item Functioning Using Iterative Hybrid Ordinal Logistic Regression/Item Response Theory and Monte Carlo Simulations [J]. Journal of Statistical Software, 2011, 39(8): 1-30.
[25] 高爽，张向葵. 应用 Rasch 模型分析 Rosenberg 自尊量表 [J]. 心理学探新, 2018, 38(05): 445-450.
[26] VAGANIAN L, BUSSMANN S, BOECKER M, et al. An item analysis according to the Rasch model of the German 12-item WHO Disability Assessment Schedule (WHODAS 2.0) [J]. Quality of Life Research, 2021, 30(10): 2929-2938.
[27] 杨治良，郝兴昌. 心理学辞典 [M]. 上海: 上海辞书出版社, 2016.
[28] COMREY A L. A first course in factor analysis. [M]. New York: Academic Press, 1973.
[29] 曹呈旭，七十三，金童林. 大学生印象管理效能感量表编制 [J]. 现代预防医学, 2021, 48(17): 3199-3201+225.
[30] 陈贵，蔡太生，胡凤姣，等. 情绪化进食量表在中国青少年中的修订 [J]. 中国临床心理学杂志, 2013, 21(4): 572-575+88.
[31] 王孟成. 潜变量建模与 Mplus 应用·基础篇 [M]. 1 ed. 重庆: 重庆大学出版社, 2013.
[32] 郭静，王瑛，宋玉萍，等. 中文版心理脆弱性问卷初步修订及在社区居民中应用信效度分析 [J]. 中国公共卫生, 2019, 35(02): 129-133.
[33] FOLYD F J, WIDAMAN K F. Factor analysis in the development and refinement of clinical assessment instruments [J]. Psychological Assessment, 1995, 7: 286-299.
[34] 刘蕾，刘华平，郭宏，等. 中文版老年人锻炼心理需求满足量表的信效度检验及适用性分析 [J]. 中国全科医学, 2021, 24(05): 619-624.
[35] 简小珠，戴步云. SPSS23.0 统计分析在心理学和教育学中的应用 [M]. 北京: 北京师范大学出版社, 2017.
[36] 刘爱梅，刘院斌. 突发性聋患者健康知信行问卷的编制和信度效度检验 [J]. 听力学及言语疾病杂志, 2012, 20(5): 444-448.
[37] 崔楚云，岳萌，李玉峰，等. 中文版卫生行为量表的信效度研究 [J]. 护理学杂志, 2017, 32(12): 62-65.
[38] 张婉玉，周艺，崔闪闪. 冠心病病人二级预防服药依从性问卷的开发与信效度检验 [J]. 护理研究, 2022, 36(6): 1004-1007.
[39] 杨丽，翟瑞龙，齐振亚，等. 心理健康素质测评系统·中国成年人认知风格问卷的编制 [J]. 心理与行为研究, 2012, 10(05): 332-339.
[40] WU Y, DONG S, LI X, et al. The Transcultural Adaptation and Validation of the Chinese Version of the Duke Anticoagulation Satisfaction Scale [J]. Frontiers in Pharmacology, 2022, 13.
[41] 游永恒，于少萍，梁斌. Beck 抑郁问卷在灾区教师中的试用及评价 [J]. 四川师范大学学报(自然科学版), 2011, 34(03): 439-442.
[42] 晏子. 心理科学领域内的客观测量——Rasch 模型之特点及发展趋势 [J]. 心理科学进展, 2010, 18(8): 1298-1305.

[43] 陈圆圆, 杨春军, 王冬梅, 等. 营养素养评价工具的汉化及在糖尿病患者中的信效度研究——基于 CTT 和 Rasch 模型的分析 [J]. 中国全科医学, 2020, 23(26): 3342-3347.
[44] 赵守盈, 何妃霞, 刘妍. Rasch 模型在学绩测验质量分析中的应用 [J]. 教育研究与实验, 2013, (01): 87-91.
[45] BAGOZZI R P, YI Y. On the evaluation of structural equation models [J]. Journal of the Academy of Marketing Science, 1988, 16(1): 74-94.
[46] 张立华, 顾莺, 黄苗, 等. 循证实践准备度评估量表的验证性因素分析 [J]. 解放军护理杂志, 2019, 36(02): 6-10+25.